

Citation for published version:

Loza Reyes, E, Hurn, MA & Robinson, A 2014, 'Classification of molecular sequence data using Bayesian phylogenetic mixture models', *Computational Statistics & Data Analysis*, vol. 75, pp. 81-95.
<https://doi.org/10.1016/j.csda.2014.01.008>

DOI:

[10.1016/j.csda.2014.01.008](https://doi.org/10.1016/j.csda.2014.01.008)

Publication date:

2014

Document Version

Peer reviewed version

[Link to publication](#)

Publisher Rights

CC BY-NC-ND

Published version available via: <http://dx.doi.org/10.1016/j.csda.2014.01.008>

University of Bath

Alternative formats

If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Classification of molecular sequence data using Bayesian phylogenetic mixture models

Elisa Loza-Reyes*, Merrilee Hurn[†] and Tony Robinson[‡]

May 23, 2013

Abstract

Rate variation among the sites of a molecular sequence is commonly found in applications of phylogenetic inference. Several approaches exist to account for this feature but they do not usually enable the investigator to pinpoint the sites that evolve under one or another rate of evolution in a straightforward manner. The focus is on Bayesian phylogenetic mixture models, augmented with allocation variables, as tools for site classification and quantification of classification uncertainty. The method does not rely on prior knowledge of site membership to classes or even the number of classes. Furthermore, it does not require correlated sites to be next to one another in the sequence alignment, unlike some phylogenetic hidden Markov or change-point models. In the approach presented, model selection on the number and type of mixture components is conducted ahead of both model estimation and site classification; the steppingstone sampler (SS) is used to select amongst competing mixture models. Example applications of simulated data and mitochondrial DNA of primates illustrate site classification via ‘augmented’ Bayesian phylogenetic mixtures. In both examples, all mixtures outperform commonly-used models of among-site rate variation and models that do not account for rate heterogeneity. The examples further demonstrate how site classification is readily available from the analysis output. The method is directly relevant to the choice of partitions in Bayesian phylogenetics, and its application may lead to the discovery

*Department of Computational and Systems Biology, Rothamsted Research, Harpenden, AL5 2JQ, UK. Email: elisa.loza@rothamsted.ac.uk

[†]Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK.

[‡]Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK.

of structure not otherwise recognised in a molecular sequence alignment. Computational aspects of Bayesian phylogenetic model estimation are discussed, including the use of simple Markov chain Monte Carlo (MCMC) moves that mix efficiently without tempering the chains. The contribution to the field of Bayesian phylogenetics is in (1) the use of mixture models augmented with allocation variables as tools for site classification and quantification of classification uncertainty, (2) the successful application of SS for selection of phylogenetic mixtures, and (3) the development of novel MCMC aspects of relevance to Bayesian phylogenetic models - whether mixtures or not¹. **Key words:** among-site rate variation; Bayesian mixture model; classification; Markov chain Monte Carlo; model selection; phylogeny.

1 Introduction

Molecular phylogenetics is the inference and interpretation of evolutionary relations between taxa based on the taxa's DNA or protein sequences. The sequences are arranged on top of one other to form an alignment with as many rows as sequences observed, and roughly as many columns (or *sites*) as characters in the sequences. The conventional likelihood-based model for phylogeny inference (e.g. Felsenstein, 1981) contains three parameters of inferential interest: a tree graph which represents the evolutionary relations between the taxa; the branch lengths of this tree which measure the expected number of nucleotide substitutions per site; and a stochastic process which models the evolution of the sequences along the branches of the tree (the latter is usually referred to as the *evolutionary model*). Such a model is complex but may still be too simple to capture important features of the generating process. In particular, it is not uncommon for sites under different functional constraints to accumulate substitutions at different rates. It is now well understood that if rate variation among sites is present and is not accounted for by the model, then spurious parameter estimates can be produced (Huelsenbeck and Suchard, 2007 and references therein).

Various approaches have been proposed to account for among-site rate variation in phylogenetic inference, including the gamma model (Yang, 1993; 1994) and several more recent models involving finite mixtures of distributions (e.g. Pagel and Meade, 2004; Lartillot and Philippe,

¹The MCMC methods discussed in this paper have been coded in a C program; source files are available upon request. Supplementary material is available online.

2004; Huelsenbeck and Suchard, 2007; Webb, Hancock and Holmes, 2009; Evans and Sullivan, 2012). The latter type of models assume that a site is generated from a mixture of multiple processes, each of which may be indexed by a specific tree topology, a specific set of branch lengths and specific parameters of the stochastic evolutionary model.

Rate variation among sites may be related to quantitative differences in the rates of substitution (e.g. sites with high rates versus sites with low rates) but also to qualitative differences in the pattern of substitution (e.g. sites with large transition/transversion rate ratios versus sites for which all substitution types occur at the same rate; Pagel and Meade, 2004). In phylogenetic applications it is possible to find quantitative among-site rate variation, qualitative variation, both or neither.

Developments in phylogenetic mixture modelling have accounted for both types of rate variation and examples of this include Felsenstein and Churchill's approach (1996). They account for quantitative variation in substitution rates among sites by a hidden Markov process that operates along the alignment, assigning rates to sites from a finite pool of values. This method incorporates the biologically realistic assumption of correlation between the rates of evolution at consecutive sites, so that the chance of neighbouring sites evolving under the same rate is higher than that of distant sites. A disadvantage of this assumption, however, is that possible biases may be introduced by the removal of sites involving gaps in the alignment, or by other errors that result in consecutive observable sites not being direct neighbours in reality.

To model qualitative rate heterogeneity, Pagel and Meade (2004) use a Bayesian mixture of multiple stochastic evolutionary processes. Their model supposes that data at a given site arise from a mixture of multiple classes, each class indexed by a common-to-all-class tree and branch lengths, and a class-specific evolutionary model. Pagel and Meade's mixture assumes a common parametrisation of evolutionary models across components. The assumption of a common set of branch lengths across mixture components results in a phylogeny whose branches are a compromise over the possibly quite different substitutional tempos in the alignment; fast and slow substitution processes are forced into a common medium. This may miss important substitutional heterogeneity and so Pagel and Meade (2008) and Meade and Pagel (2008) consider extensions to their original model, this time allowing for multiple sets of branch lengths.

Kolaczkowski and Thornton (2008) present a mixture similar to that in Meade and Pagel (2008), but conduct inference within a maximum-likelihood framework.

Evans and Sullivan (2012) model both quantitative and qualitative rate heterogeneity using mixtures whose components share a common tree topology and set of branch lengths but are indexed by individual evolutionary models, each accompanied by a scaling factor that permits each mixture component to follow its own tempo of substitution. They further allow for different substitution rate constraint cases of the evolutionary GTR model across components and conduct inference on both GTR constraint cases and the number of mixture components via reversible jump MCMC.

A related approach, called the CAT model (Lartillot and Philippe, 2004), considers qualitative mixtures of stochastic evolutionary processes, all with the same set of substitution rates but with different stationary probabilities. Inference on the number of mixture components is conducted using a Dirichlet process prior and the model is estimated via MCMC. Huelsenbeck and Suchard (2007) consider quantitative mixtures of branch lengths in which sites are partitioned into classes according to a Dirichlet process prior. Sites that are assigned to the same class share a common set of branch lengths, while all sites, irrespective of their class, share a common topology and evolutionary model. Both the number of classes and the assignment of sites to classes are treated as random variables and, together with the usual phylogenetic parameters, are objects of inferential interest.

One aspect of mixture models that has been under-explored in the phylogenetics literature is their use for site classification through the introduction of latent allocation variables. The allocation variables identify the underlying class of a site and thus enable the decomposition of the complicated structure of a mixture into simpler structures. In a phylogenetic context, mixture components may have a direct biological interpretation and site classification can lead to insights of structure and heterogeneity in the alignment that are not otherwise easily uncovered. The purpose of this study is, therefore, to extend the functionality of phylogenetic mixtures to include allocation variables and investigate their use for site classification. The mixture in Pagel and Meade (2004) is a pioneering example of the use of phylogenetic mixtures for site classification. Their model does not incorporate allocation variables and site classification involves a

posteriori processing of the analysis output to determine the component that is most likely to have generated a site (e.g. recent work by Xi et al., 2012 applies Pagel and Meade’s classification approach to interrogate angiosperm molecular sequence data and thus to improve phylogenetic resolution in this group of plants). Lartillot and Philippe (2004) and Huelsenbeck and Suchard (2007) incorporate allocation variables, but straightforward statements about site classification are obscured by their consideration of the number of mixture components as a random variable. Their MCMC samplers move between mixtures of different dimensions rendering the allocation variables useless for straightforward classification (Lartillot and Philippe, 2004, resort to a *posteriori* multivariate analysis techniques to identify stable components across their MCMC run). More recently, Evans and Sullivan (2012) do not attempt site classification but acknowledge its importance as a means to extract structural information from the observed data and call it “a very hard problem made even harder by the estimation of the number of mixture components”.

The approach pursued in this study is to model the class allocation of a site jointly with the rest of the phylogenetic mixture parameters. We do not consider the number of mixture components as a random variable and, in contrast with previous published approaches, determine this parameter using model selection techniques ahead of model estimation. In the absence of label switching, the output from the model estimation stage can be directly used for site classification. We employ the steppingstone sampler (SS) of Xie et al. (2011) to estimate the marginal likelihood of a model and thus conduct model selection. We demonstrate that SS is able to correctly recover the true model when applied to simulated data, and further apply it to analyse mitochondrial DNA data. We perform site classification using the model selected by SS, and demonstrate that our use of phylogenetic mixtures augmented by allocation variables correctly detects heterogeneity in the data and accurately classifies the sites to evolutionary components. This approach estimates the posterior probability distribution of site allocation so that classification uncertainty may also be appreciated. As part of our MCMC implementation, we present a novel set of move types to update the parameters of a mixture phylogenetic model, and investigate their performance. We show that our MCMC algorithm achieves the same, or greater, efficiency than existing methods with potential for deployment in the estimation of both mixture or non-mixture phylogenetic models at reduced computational cost.

2 Bayesian phylogenetic mixtures for site classification

2.1 The models

The backbone of likelihood-based phylogenetic methods is a *homogeneous model* positing that the characters at a site in a DNA alignment are an independent realisation of a continuous-time Markov process, with state space $\mathcal{I} = \{A, C, G, T\}$, that evolves on the branches of a bifurcating tree topology, ϕ , and has realisations at the leaves of this tree. The instantaneous rate matrix, Q , that generates the Markov process is indexed by a (possibly vector) parameter θ . There are several proposed parametrisations of the Q -matrix in the literature (e.g. Jukes and Cantor, 1969; Hasegawa et al., 1985) with the most general time-reversible one called the GTR matrix (Lanave et al., 1984; Tavaré, 1986), where

$$Q(\theta) = \begin{pmatrix} q_{AA} & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ r_{AC}\pi_A & q_{CC} & r_{CG}\pi_G & r_{CT}\pi_T \\ r_{AG}\pi_A & r_{CG}\pi_C & q_{GG} & r_{GT}\pi_T \\ r_{AT}\pi_A & r_{CT}\pi_C & r_{GT}\pi_G & q_{TT} \end{pmatrix} \quad (1)$$

and $\theta = (r, \pi)$ is a collection of six substitution rates $r = (r_{AC}, \dots, r_{GT})$ and four stationary probabilities $\pi = (\pi_A, \dots, \pi_T)$ with constraints $r_m, \pi_i \geq 0$ and $\sum r_m = \sum \pi_i = 1$ ($m = AC, AG, \dots, CT, GT; i = A, C, G, T$). The diagonal values of Q are defined so that each row adds up to zero. The expected total rate of substitution of the process generated by matrix Q is equalled to one (Felsenstein, 1981) so that the branch lengths represent the expected number of substitutions per site. The Markov process of character substitution is time-reversible, a feature that prevents us from inferring rooted trees. Thus, for an observed alignment of size S sequences $\times N$ sites, parameter ϕ takes values in the set of unrooted bifurcating leaf-labelled trees for S taxa; branch lengths are real valued; and the space in which parameter θ takes values is dictated by the chosen parametrisation of matrix Q . The objective of the analysis is usually inference about the tree topology, ϕ , this tree's branch lengths (denoted by a set $t = \{t_1, \dots, t_{2S-3}\}$) and θ .

Building upon the homogeneous model, we account for among-site rate variation using a

finite mixture of distributions of the type

$$x_n \mid \omega, \phi, t, \theta, k \sim \sum_{j=1}^k \omega_j p(x_n \mid \phi, t_j, \theta_j), \text{ independently for } n = 1, \dots, N, \quad (2)$$

where x_n is the observed data at site n ; k is the number of mixture components; $\omega = (\omega_1, \dots, \omega_k)$ are the mixture proportions ($\omega_j \geq 0$ and $\sum_{j=1}^k \omega_j = 1$); each component j ($j = 1, \dots, k$) has set of branch lengths t_j and parameters of the Q -matrix θ_j collectively denoted by $t = (t_1, \dots, t_k)$ and $\theta = (\theta_1, \dots, \theta_k)$; and $p(x_n \mid \phi, t_1, \theta_1), \dots, p(x_n \mid \phi, t_k, \theta_k)$ are the k component likelihoods. Model (2) thus asserts that characters at site n are generated from a mixture of k different evolutionary components occurring in proportions $\omega_1, \dots, \omega_k$. To decompose the structure of this mixture, we extend the model to include a set of N scalar latent allocation variables, $z = (z_1, \dots, z_N)$, where each $z_n \in \{1, \dots, k\}$. Extending Bayesian mixture models to include allocation variables is fairly standard in the statistics literature (e.g. Richardson and Green, 1997), sometimes for their own sake when classification is the goal and other times just because they can simplify the computational fitting. The n^{th} of these latent variables answers the question “to which component does observation n belong?”. Including z in the model, we can write

$$x_n \mid \phi, t, \theta, k, z_n \sim p(x_n \mid \phi, t_{z_n}, \theta_{z_n}), \text{ independently for } n = 1, \dots, N. \quad (3)$$

This formulation not only accounts for both quantitative and qualitative rate heterogeneity, but also provides a means to class discovery by the use of z . In addition to classifying the sites to evolutionary components, mixture (3) also enables us to discern the profile of each class by estimating the component-specific parameters. So, the analysis may lead to statements such as “class 1 is more conserved than class 2 as the former displays a shorter total branch length than the latter” or “the nucleotide composition of the two classes is quite different, as reflected by the estimated stationary probabilities”.

The joint prior of all parameters is expressed as

$$\begin{aligned}
p(\omega, z, \phi, \theta, t) &= p(\omega)p(z | \omega)p(\phi | z, \omega)p(\theta | \phi, z, \omega)p(t | \theta, \phi, z, \omega) \\
&= p(\omega)p(z | \omega)p(\phi)p(t)p(\theta)
\end{aligned} \tag{4}$$

where we have suppressed the explicit conditioning on k because we consider only mixtures with a fixed number of components and make independence assumptions between all parameters other than z and ω . The prior for ω is taken to be the symmetric Dirichlet distribution $\omega \sim \text{Dir}_k(\rho, \dots, \rho)$. We express prior ignorance about class size by setting $\rho = 1$.

Conditional on ω , the allocations z_1, \dots, z_N are assumed independent and identically distributed

$$\Pr(z_n = j | \omega) = \omega_j, \quad j = 1, \dots, k. \tag{5}$$

We make the following standard choices for the priors on phylogenetic parameters. All tree topologies are assumed to be equally likely *a priori*; that is, we take a discrete uniform prior for ϕ (e.g. Suchard et al., 2001). The prior distribution for branch lengths makes an assumption that the $2S - 3$ branches for each of the k components are independent both within components and across components. Exponential priors on individual branch lengths are specified, with exponential-rate parameter η so that $\mathbf{E}(t_{h,j}) = 1/\eta$ for branch length h in the j th mixture component ($h = 1, \dots, 2S - 3; j = 1, \dots, k$). For the parameter vectors θ of the k instantaneous rate matrices, we assume independent prior distributions on each r_j and π_j of the form $r_j \sim \text{Dir}_6(1, \dots, 1)$ and $\pi_j \sim \text{Dir}_4(1, \dots, 1)$.

Throughout, the model specified by (2) and (3) is referred to as the $Q + t$ mixture model.

We also consider nested submodels of the $Q + t$ mixture. Firstly, we consider mixtures of multiple Q matrices which share a common set of branch lengths, t , and tree topology, ϕ , (Pagel and Meade, 2004):

$$x_n | \omega, \phi, t, \theta, k \sim \sum_{j=1}^k \omega_j p(x_n | \phi, t, \theta_j), \quad \text{independently for } n = 1, \dots, N. \tag{6}$$

Restricting this further we consider mixtures of branch lengths and Q matrices, but where the Q matrices across components share the same stationary probabilities, i.e. $\theta_1 = (r_1, \pi), \dots, \theta_k =$

(r_k, π) :

$$x_n \mid \omega, \phi, t, \theta, k \sim \sum_{j=1}^k \omega_j p(x_n \mid \phi, t_j, r_j, \pi), \text{ independently for } n = 1, \dots, N. \quad (7)$$

Both models can be augmented with allocation variables. We refer to model (6) and its corresponding augmented formulation as the Q *mixture*, and to model (7) and its augmented version as the $r + t$ *mixture*. For simplicity, we use the notation $Q + t(k)$ to denote a $Q + t$ mixture with k components, and similarly for $Q(k)$ and $r + t(k)$.

Our mixture models share certain similarities with those used in Pagel and Meade (2004, 2008), Meade and Pagel (2008), Lartillot and Philippe (2004), Huelsenbeck and Suchard (2007) and Evans and Sullivan (2012) but differ in one key way. They are formulated to include allocation variables and the posterior probability of these variables is an object of primary inferential interest and a means for both site classification and quantification of classification uncertainty.

2.2 Likelihood computation

The likelihood function under the most general $Q + t$ mixture is the product of the distributions at individual sites (equation (3)), from site 1 to N :

$$L(\phi, t, \theta \mid x, z) = \prod_{n=1}^N p(x_n \mid \phi, t_{z_n}, \theta_{z_n}). \quad (8)$$

We assume that substitutions at different branches of the tree and among different sites in the alignment are independent of one another. Likelihood (8) is usually computed for a specific tree and so each tree topology requires a reformulation of this function according to its corresponding branching structure; the larger the tree the more computationally prohibitive the calculation. A recursive technique for the efficient computation of phylogenetic likelihood functions, called the pruning algorithm, was introduced by Felsenstein (1981), and this is the algorithm that we use.

3 Model estimation

Markov chain Monte Carlo (MCMC) will be required to fit models of this complexity and we present the basic move types in our MCMC sampler. A distinctive feature of our method is that changes to the topology are separated from those in branch lengths; this is particularly important for some of the mixtures where the components share a common topology but have different sets of branch lengths. Metropolis-Hastings methods are equally valid when the available moves are scanned either randomly or systematically. Here, we have chosen to take the latter approach making use of six move types:

- (a) updating the tree topology ϕ ;
- (b) updating all branch lengths $t_1, t_2, \dots, t_{2S-3}$;
- (c) updating the vector of substitution rates r ;
- (d) updating the vector of stationary probabilities π ;
- (e) updating the vector of mixture proportions ω ;
- (f) updating all site allocations z_1, z_2, \dots, z_N .

Moves (b) – (d) are applied to all components in the mixture, if required. One complete pass over these six moves is an iteration, the basic time step of our MCMC sampler. The first two move types focus on the tree while the next two concentrate on the parameters of the models on the tree; the last two move types concern the mixture allocations and proportions. We now consider the three groups separately in the context of the most general $Q + t$ mixture model.

3.1 Updating the tree topology and branch lengths

The tree topology is updated via the nearest neighbour interchange (NNI) (Robinson, 1971; Moore, Goodman and Barnabas, 1973), in which one of the two nearest neighbours of the current topology (in NNI space) is proposed with equal probability. NNI generates a candidate topology while preserving the current set of branch lengths. A candidate topology ϕ' is accepted with a probability that simplifies to:

$$a(\phi, \phi') = \min \left\{ 1, \frac{L(\phi', t, \theta \mid x, z)}{L(\phi, t, \theta \mid x, z)} \right\}. \quad (9)$$

A separate proposal mechanism is used to update the branch lengths while maintaining the same topology. We consider two different proposals for branch lengths:

- Branch length multiplier (BLM). Also known as proportional shrinking and expanding (Yang, 2006), this proposal updates the length of a randomly chosen branch $t_{h,j}$ by multiplying it by a quantity m generated from the density

$$f(m) = (\lambda m)^{-1}, \quad 1/\delta < m < \delta \quad (10)$$

where $\lambda = 2 \log \delta$ and $\delta > 1$ acts as a tuning parameter.

- Branch length normal additive (BLNA). Also known as the sliding window proposal (Huelsenbeck and Ronquist, 2001), this mechanism updates a randomly chosen branch length $t_{h,j}$ via an additive Gaussian perturbation, $t'_{h,j} \sim N(t_{h,j}, \sigma^2)$, so that σ^2 acts as the tuning parameter. If negative branch lengths are proposed, they are reflected at zero with the proposal still remaining symmetric.

BLM may be thought of as self-tuning as the variance of the proposed branch length is proportional to the square of the original length. This works well when exploring large branches but can be a bit sticky when branch lengths are small as it can take a large number of iterations to move a short distance. On the other hand, a candidate branch length generated from the BLNA proposal has a step size which depends only on the tuning parameter σ^2 and not on the current branch length. This makes it hard for BLNA to work equally effectively at both large and small scales. In experiments, we achieved best performance by alternating between BLM and a BLNA tuned for small branch lengths (Supplementary Material I).

The acceptance probability of a branch length proposed from either BLM or BLNA is

$$a(t_{h,j}, t'_{h,j}) = \min \left\{ 1, \frac{p(t'_{h,j})}{p(t_{h,j})} \frac{L(\phi, t', \theta \mid x, z)}{L(\phi, t, \theta \mid x, z)} \frac{q(t'_{h,j}, t_{h,j})}{q(t_{h,j}, t'_{h,j})} \right\}. \quad (11)$$

The proposal ratio $q(t'_{h,j}, t_{h,j})/q(t_{h,j}, t'_{h,j})$ simplifies to m for BLM and to 1 for BLNA, and so acceptance $a(t_{h,j}, t'_{h,j})$ simplifies to $m e^{-\eta(t'_{h,j} - t_{h,j})}$ and to $e^{-\eta(t'_{h,j} - t_{h,j})}$ for BLM and BLNA, respectively, times the likelihood ratio in both cases.

3.2 Updating the Markov process parameters

The j th component of the $Q + t$ mixture has a set of parameters controlling the substitution rates plus a set of stationary probabilities, $r_{AC,j}, \dots, r_{GT,j}$ and $\pi_{A,j}, \dots, \pi_{T,j}$, respectively. Since we can treat each mixture component separately for updating purposes, we drop the subscript j . Both types of parameters are constrained to sum to one and, as they utilise the same type of proposal, here we concentrate on the substitution rates.

We generate a new set of substitution rates, r' , from a Dirichlet distribution centred at the current rate values with a positive shift $\epsilon > 0$ and with tuning parameter $\alpha > 0$; i.e. $r' \sim \text{Dir}_6(\alpha(r_{AC} + \epsilon), \dots, \alpha(r_{GT} + \epsilon))$. The variance of the m th element of a rate vector proposed with this move type (henceforth referred to as the ϵ Dirichlet proposal) is:

$$\text{var}(r'_m) = \frac{\alpha(r_m + \epsilon)(\alpha_0 - \alpha(r_m + \epsilon))}{\alpha_0^2(\alpha_0 + 1)} \quad (12)$$

where $\alpha_0 = \sum_{m=AC}^{GT} \alpha(r_m + \epsilon) = \alpha(1 + 6\epsilon)$. When $\epsilon = 0$, our proposal becomes the popular Dirichlet proposal by Larget and Simon (1999), which suffers from one major drawback: when r_m is close to zero so too is $\text{var}(r'_m)$. This can create an undesirable cycle in which the MCMC sampler keeps proposing candidate rates very close to zero because the step size of the proposal is nearly zero, typically needing many iterations to escape. We introduce the offset ϵ as an effective way to improve the mixing of the chain without resorting to tempered schemes that can result in high computational burden. The effectiveness of this proposal is investigated in Supplementary Material II. The move is accepted with probability

$$a(r, r') = \min \left\{ 1, \frac{L(\phi, t, \theta' \mid x, z)}{L(\phi, t, \theta \mid x, z)} \frac{q(r', r)}{q(r, r')} \right\} \quad (13)$$

where the proposal ratio $q(r', r)/q(r, r')$ is calculated as the quotient of two Dirichlet density functions.

3.3 Updating the mixture parameters

Updating the allocation variables and the vector of mixture proportions is a fairly standard problem in the estimation of Bayesian mixtures via MCMC. The vector of mixture proportions

$\omega = (\omega_1, \dots, \omega_k)$ is usually updated using a Gibbs sampler since their posterior conditional is easily seen to be a Dirichlet distribution with parameters $\rho + N_1, \dots, \rho + N_k$, where $N_j = \sum_{n=1}^N I[z_n = j]$ is the number of sites allocated to component j and $I[\cdot]$ is the indicator function. This mechanism thus updates ω according to the number of sites allocated to each component on a given MCMC iteration. A well known difficulty of this proposal is that it may mix badly when one or more components become quite small or when the other parameters characterising the components make it hard for a site to swap components (see Leslie, 2007 or Hurn et al., 2008 for examples in quite different application areas). In the latter case, Leslie (2007) and Hurn et al. (2008) both suggest a strategy that updates ω and the allocations jointly. However here we are primarily worried about instances in the MCMC path when one or more components become quite small, causing the chain to mix badly. Given our experience in updating r and π , we use a shifted Dirichlet approach, here replacing the Gibbs draw from a $Dir_k(\rho + N_1, \dots, \rho + N_k)$ by a Metropolis-Hastings proposal, $\omega' \sim Dir_k(\rho + N_1 + \epsilon, \dots, \rho + N_k + \epsilon)$ with $\epsilon > 0$. The acceptance probability of this move type simplifies to

$$a(\omega, \omega') = \min \left\{ 1, \prod_{j=1}^k (\omega_j / \omega'_j)^\epsilon \right\} \quad (14)$$

and so a high acceptance rate is maintained for small values of ϵ .

The allocation for the n th site, z_n , is updated by drawing, randomly and with equal probability, from the set $\{1, \dots, k\}_{-z_n}$ where the subindex denotes that z_n is excluded from the set. Since allocations are updated one at a time, the acceptance probability involves a ratio of likelihoods only at site n :

$$a(z_n, z'_n) = \min \left\{ 1, \frac{\omega_{z'_n}}{\omega_{z_n}} \frac{p(x_n | \phi, t_{z'_n}, \theta_{z'_n})}{p(x_n | \phi, t_{z_n}, \theta_{z_n})} \right\}. \quad (15)$$

3.4 The label switching problem

The application of MCMC methods to mixture models suffers from the problem that the components of a mixture are intrinsically non-identifiable since the same likelihood function is obtained for any permutation of the component labels. If the prior distributions do not contain information regarding labelling of the components, this may produce MCMC output that switches

labelling throughout the run and masks component-specific information. In that case, label switching needs to be dealt with before any meaningful inference can be made on parameters for individual mixture components. Our classification method relies on component-specific inferences and so, if label switching is detected, it is necessary to deal with it. A description of this phenomenon, together with a review of existing solutions, can be found in Jasra et al., (2005). We note that the commonly used approach of imposing an ordering on one of the parameters, used for example in Richardson and Green (1997), can be applied after running the simulations when it is clearer which parameter is most effective to use for unpicking the label switching for inference purposes.

4 Model selection via steppingstone sampling

We now turn to the choice of which model to use for a particular set of data. Bayes factors (BFs) can be computed to summarise the evidence provided by the data in favour of one model relative to another (Kass and Raftery, 1995). When two models are equally likely *a priori*, the BF is defined as the ratio of the marginal likelihood under model M_1 to the marginal likelihood under a second model, M_0 , given the data, x . BFs are usually interpreted on the log scale using the rule of thumb that $2\ln(\text{BF}) > 10$ indicates very strong evidence in favour of model M_1 , $0 \leq 2\ln(\text{BF}) \leq 2$ indicates no significant difference between the models, and with a range of levels in between according to a scale provided in Kass and Raftery (1995).

The Savage-Dickey ratio (Verdinelli and Wasserman 1995; Suchard et al. 2001) has been successfully used in phylogenetics to estimate BFs directly. This approach requires that models are nested and, although potentially useful for some models in our problem, we have not used this approach here. There exist a number of other ways to estimate BFs; Friel and Pettitt (2008) categorise these broadly into *across model* or *within model* approaches. The most common examples of the former type use reversible jump MCMC (RJMCMC; Green, 1995), embedding the entire set of possible models within a larger model and then sampling from this joint posterior of both models and parameters. Under this approach, BFs of pairs of the competing models are estimated by the relative proportions of time the sampler spends in the two models. Within model approaches, on the other hand, consider the competing models one by one, estimating each

marginal likelihood in turn in order to estimate the BFs. Friel and Pettitt (2008) suggest that the across model approach is the better suited of the two in situations where there are many competing models. Set against this though is the well documented difficulty in constructing efficient RJMCMC algorithms, combined with the associated long run times required to achieve stable estimates. This is particularly a problem when the competing models do not relate to one other in a way that suggests natural moves between their sets of parameters.

We have opted to follow a within model approach and so need to estimate the marginal likelihood for each competing model. The marginal likelihood for model M_i is the expectation (under the prior) of the likelihood of the data x , conditioned on the model M_i (or, equivalently, the integral over the parameters of the joint distribution of the data and the prior conditioned on the model),

$$p(x | M_i) = \int_{\vartheta_i} p(x | \vartheta_i, M_i) p(\vartheta_i | M_i) d\vartheta_i \quad (16)$$

where ϑ_i is the parameter vector of model M_i . The marginal likelihood in equation (16) cannot be calculated analytically except for the most elementary applications, and its estimation is the topic of considerable interest (e.g. recent work by Ardia et al. (2012) reviews popular Monte Carlo methods for marginal likelihood estimation). An unfortunately commonly used estimator is the harmonic mean (HM) of Newton and Raftery (1994) which is a form of importance sampling, taking the posterior as its importance distribution. HM marginal likelihoods can be calculated from the MCMC chain used for fitting the model at little extra cost but generally perform poorly. Two alternatives that have been used in the phylogenetic literature are thermodynamic integration (Lartillot and Philippe, 2006; Friel and Pettitt, 2008) and steppingstone sampling (SS) in either its original (Xie et al., 2011) or generalised (Fan et al., 2011) flavour.

Marginal likelihood estimation via the original SS method has been shown to be more accurate than both thermodynamic integration and HM in applications to Bayesian phylogenetics (Xie et al., 2011; Baele et al., 2012), while the generalised flavour of SS improves upon its original version in terms of efficiency and stability. Nevertheless, generalised SS requires the specification of a reference distribution that approximates the posterior of interest; for complex phylogenetic mixture models it is unclear how such a reference can be chosen.

The original SS method (henceforth referred to simply as SS) estimates the marginal likelihood from several MCMC runs stitched together along a path that goes from the posterior to the prior with differing power posterior distributions in between. The power posterior distribution under model M_i and with parameter vector ϑ_i is:

$$p_\beta(\vartheta_i | x) = \frac{p(x | \vartheta_i)^\beta p(\vartheta_i)}{c_\beta} \quad (17)$$

where $0 \leq \beta \leq 1$ and c_β is a normalising constant. Dependence on the model under consideration (M_i) has been suppressed in the notation for simplicity. When $\beta = 1$, the power posterior is the posterior distribution and the normalising constant c_1 is the marginal likelihood, i.e. $c_1 = p(x)$. The power posterior is equivalent to the prior distribution when $\beta = 0$ and, assuming that the prior is proper, $c_0 = 1$. The basic idea of SS is to express the marginal likelihood as the product of K ratios:

$$\begin{aligned} p(x) &= \frac{c_1}{c_0} \\ &= \left(\frac{c_{\beta_1}}{c_{\beta_0}} \right) \dots \left(\frac{c_{\beta_\nu}}{c_{\beta_{\nu-1}}} \right) \dots \left(\frac{c_{\beta_K}}{c_{\beta_{K-1}}} \right). \end{aligned} \quad (18)$$

where $0 = \beta_0 < \dots < \beta_{\nu-1} < \beta_\nu < \dots < \beta_K = 1$ are the stepping stones between the prior and the posterior distributions. Each ratio $c_{\beta_\nu}/c_{\beta_{\nu-1}}$ is estimated as the average value of the observed likelihoods raised to the power $\beta_\nu - \beta_{\nu-1}$ when sampled from a MCMC run with target distribution $p_{\beta_{\nu-1}}$. Therefore, SS does not require samples from the posterior. In practice, however, we start by sampling from the posterior to burn-in the chain and proceed in the direction $p_{\beta_{K-1}}$ until reaching the prior p_{β_0} .

To construct the SS sampler from our MCMC algorithm, the proposals remained unchanged and the likelihood ratios in acceptance probabilities $a(\phi, \phi')$, $a(t_{h,j}, t'_{h,j})$, $a(r, r')$ and $a(z_n, z'_n)$ were raised to the power $\beta_{\nu-1}$. We set $K = 30$ and spaced the values of β according to uniform quantiles of a $Beta(0.3, 1)$ distribution which, in practice, entailed setting $\beta_\nu = (\nu/K)^{3.33}$. Xie et al. (2011) show that the accuracy of SS is optimal in a Gaussian model example when β values are set in this way. Assessing optimal specification of β in Bayesian phylogenetic mixture applications is outside the scope of this study so we followed the Xie et al. (2011)

recommendation. However, see also some recent work by Friel, Hurn and Wyse (2013).

We employed SS to select between mixture models with differing parameterisations (e.g. $Q + t$ versus Q) and also to assess the number of components in a given mixture (e.g. $k = 2$ versus $k = 3$). The aim of model selection is not necessarily to find the true model that generated the data but to select a model that captures the key features of the data while being biologically realistic and tractable (Steel, 2005). Only once we have determined the most plausible mixture and its number of component subpopulations for a given data set, do we perform site classification.

5 Classification of simulated data

5.1 Methods

To validate our classification approach, we generated a synthetic DNA alignment of size 16 sequences \times 2500 sites, with the software package Seq-Gen (Rambaut and Grassly, 1997). Sites 1 – 1500 were generated from an evolutionary class with substitution rates $\{r_{AC} = r_{AT} = r_{CG} = r_{GT} = 0.0500, r_{AG} = r_{CT} = 0.4000\}$, stationary probabilities $\{\pi_A = 0.3220, \pi_C = 0.3040, \pi_G = 0.1080, \pi_T = 0.2660\}$ and total branch length $T = 10$. Sites 1501 – 2500 were simulated with $\{r_{AC} = 0.1009, r_{AG} = 0.3645, r_{AT} = 0.1506, r_{CG} = 0.0639, r_{CT} = 0.3044, r_{GT} = 0.0157\}$; $\{\pi_A = \dots = \pi_T = 0.2500\}$ and $T = 0.1$. Both classes were generated under the same tree topology, which was randomly sampled from the space of all unrooted bifurcating trees that relate 16 sequences. In our experiments, the topology was held fixed at its generating value.

The intention here is to assess whether the classification method is able to detect the substitutional differences between the two classes and to correctly allocate sites to evolutionary groups without prior knowledge of the partitioning in the data.

5.2 Model selection

Before the runs for inference, we conducted several exploratory runs to tune the SS proposals to $\delta = 1.5$ for the BLM move; $\sigma = 0.08$ for BLNA; $\alpha_r = 900$; $\alpha_\pi = 700$ and $\epsilon_\theta = 0.0001$ for ϵ Dirichlet of substitution rates and stationary probabilities; and $\epsilon_\omega = 0.0001$ for ϵ Dirichlet of

mixture proportions. Hyperparameter η for the prior on a branch length was set to 4.5 as a compromise between the two simulated classes. The SS sampler was run for 5 000 iterations for each β value in the steppingstone path. The burn-in phase consisted of 20 000 iterations at power $\beta = 1$ (the posterior). It is worth noting that one iteration in our MCMC sampler systematically updates all the parameters in the model. So, there are $2 + k(2S - 1) + N$ parameter updates per iteration in a fit of a $Q + t$ mixture with k components to an alignment of S sequences and N sites (one update for the topology, $k(2S - 3)$ for all branch lengths across all mixture components, $2k$ for the rate and stationary probability vectors across all components, N for all site allocations and one for the vector of mixture proportions). Care must be therefore taken when assessing the length of our runs; 5 000 iterations here correspond to 12.8×10^6 parameter updates.

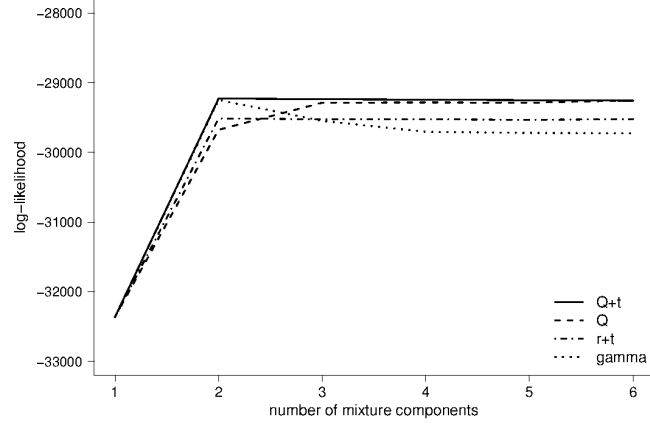
Models (2), (6) and (7), in their allocation-variable formulation, were considered for the synthetic alignment with $k = 1, \dots, 6$ components. Figure 1(a) shows the log marginal likelihoods for these models, estimated using the SS sampler. The log-likelihood for $k = 1$ is common across all mixture types and it corresponds to a fit of the data with the homogeneous model. It is clear that any mixture fits these data better than the homogeneous model, which is unsurprising given the heterogeneity that underlies the alignment. For comparison, we also estimated the log marginal likelihood of a special type of mixture model in which all components share one Q and set of branch lengths, but each component is allowed a separate scalar γ_j that scales the Q matrix and that is drawn from a discrete version of a gamma distribution with empirically-estimated shape parameter (Yang, 1993; 1994; see also Pagel and Meade, 2004 for a description in the context of mixture models). The log marginal likelihood of the discrete-gamma model, as this formulation is usually known, was estimated for 2 – 6 categories of the discrete gamma distribution using the software package MrBayes 3.2 (Ronquist et al., 2012). We specified the number of MCMC cycles and settings so that MrBayes’ analysis was comparable to our SS sampler. The discrete-gamma model is a popular choice in phylogenetics because it accounts for rate heterogeneity in an elegant and convenient way by adding only one extra parameter to the homogeneous model formulation; the shape parameter of the gamma distribution. But such convenience and elegance may be insufficient when dealing with complex evolutionary scenarios.

Figure 1(a) shows that the two best performing models are $Q + t(2)$ and a 2-category discrete gamma model. According to Kass and Raftery’s scale, the former provides a significantly better fit to the data than the latter, which suggests that the substitutional heterogeneity in the data can only be adequately explained by a mixture of Q matrices and confirms that our SS sampler is able to select the true model as the most plausible one.

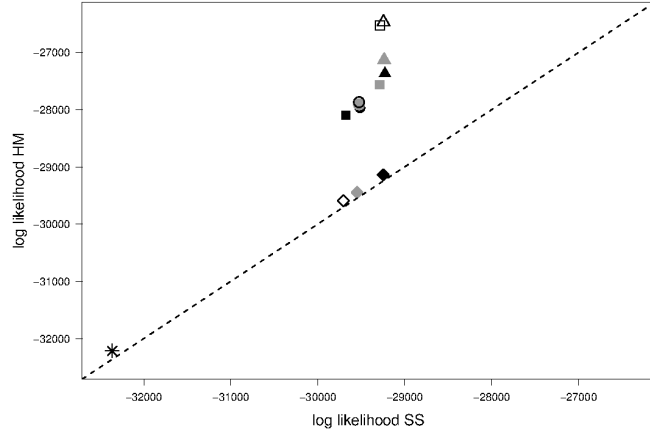
5.3 Comparison to other marginal likelihood estimators

As a comparison, we conducted model selection using the HM estimator, which can be straightforwardly calculated from the chain of log-likelihoods returned by the MCMC posterior simulation. We simulated 60 000 samples from each of the allocation-variable versions of models (2), (6) and (7) with $k = 1, \dots, 4$ components. We thinned the samples to every 10 iterations and discarded the first quarter as burn-in. For mixtures with four components, we simulated 20 000 iterations and thinned and burnt-in the chain in a similar way. We tuned the parameters as for the SS run. We also simulated from a discrete-gamma model using MrBayes with comparable settings and computational effort as for our MCMC runs for HM estimation.

In Figure 1(b) we have plotted the log-marginal-likelihoods obtained using HM against SS for all considered models. HM estimates exceeded those of SS in all cases, and the pattern is exacerbated as the models become more complex. The difference between the HM and the SS estimates for $Q + t(4)$ is as large as 2 776 log units, and almost as large for $Q(4)$. Indeed, HM selects a $Q + t(4)$ mixture even though the data was generated under the simpler $Q + t(2)$ model. The reason is that HM fails to adequately penalise the more complex models for having extra parameters that contribute little to model fit (Xie et al., 2011). These results coincide with the growing evidence that HM often overestimates the marginal likelihood making a model appear better-fitting than it really is (Lartillot and Philippe, 2006; Xie et al., 2011; Baele et al., 2012) and support the notion that HM should be avoided (Calderhead and Girolami, 2009). We do note that our HM and SS estimates are not based on the same number of samples but suggest that the observed patterns in Figure 1(b) will not be significantly influenced by this.



(a)



(b)

Figure 1: (a) Estimated log marginal likelihoods for models considered for the synthetic DNA alignment. Plotted data: homogeneous: $-32\,364$; $Q+t$ with 2–6 components: $-29\,226$, $-29\,235$, $-29\,243$, $-29\,250$, $-29\,257$; Q with 2–6 components: $-29\,675$, $-29\,287$, $-29\,282$, $-29\,288$, $-29\,253$; $r+t$ with 2–6 components: $-29\,513$, $-29\,524$, $-29\,525$, $-29\,532$, $-29\,522$; gamma with 2–6 categories: $-29\,246$, $-29\,546$, $-29\,704$, $-29\,720$, $-29\,726$. (b) Scatter plot of log marginal likelihood values estimated using HM versus SS. Plotted values on the x-axis are given in legend of Fig. 1(a). Plotted values on the y-axis (geometric symbols in black, grey and white fill correspond to 2, 3 and 4 model components/categories, respectively): homogeneous (*): $-32\,216$; $Q+t$ (\triangle) with 2–4 components: $-27\,369$, $-27\,134$, $-26\,467$; Q (\square) with 2–4 components: $-28\,099$, $-27\,563$, $-26\,530$; $r+t$ (\circ) with 2–4 components: $-27\,966$, $-27\,908$, $-27\,868$; gamma (\diamond) with 2–4 categories: $-29\,138$, $-29\,447$, $-29\,594$. Dashed line: region in which HM and SS estimates would agree.

5.4 Model estimation and classification

We estimated the $Q + t$ mixture with two components twice, and verified that each of these independent runs converged to the same region in the posterior distribution. The runs for inference comprised 60 000 iterations of our MCMC sampler thinned to every 10 iterations, and we discarded the first quarter as burn-in. The tuning parameters for the proposals remained at the same values as for the SS run. Figure 2 shows the estimated posterior probabilities of classification to the two components which, unlike previously published methods (e.g. Pagel and Meade, 2004; Lartillot and Philippe, 2004; Huelsenbeck and Suchard, 2007), can be directly obtained from the MCMC output as follows. Let $z_n^{(1)}, \dots, z_n^{(M)}$ be the chain of posterior allocations for site n , generated by an MCMC run of length M after burn-in. Variable $z_n^{(i)}$ indicates the identity of the component to which site n is allocated at iteration i and it takes values in the set $\{1, \dots, k\}$. Once the chain is checked for convergence to stationarity, good mixing and lack of label-switching, it is used to count the number of times that site n gets allocated to component j . This frequency count, divided by the total number of samples, M , gives the *posterior classification probability* of site n to component j . (Lack of label-switching can be visually verified by inspection of the chain traces; for instance, the r_{AT} and π_G traces in Supplementary Material III show that the label - or the colour in the case of our visualisation - of each component remains consistent throughout the MCMC run.) In Figure 2, the crossover at which sites were simulated from different evolutionary classes was strikingly well recovered by the method and ergodic posterior averages for the remaining parameters in the mixture coincided favourably with the generating values (Supplementary Material III). The posterior classification probabilities offer a means for not only site classification but also for quantification of classification uncertainty.

6 Classification of mitochondrial DNA

6.1 Methods

In a second application, we revisited the analysis of mitochondrial DNA (mtDNA) sequences from the primate species human; gorilla; chimpanzee; orangutan; gibbon; crab-eating macaque; common squirrel monkey; Philippine tarsier and ring-tailed lemur (Brown et al., 1982; Hayasaka

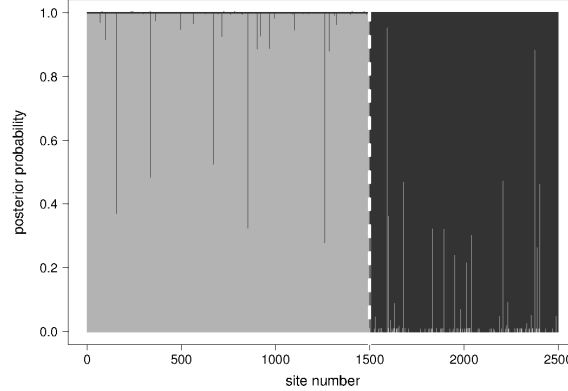


Figure 2: Posterior classification probabilities for the synthetic DNA alignment, from simulation from the posterior of a two-component $Q + t$ mixture model. ■ and ■ denote the two different mixture components. The dotted line indicates the boundary between the evolutionary classes at generation stage.

et al., 1988). This alignment, of size 9 sequences \times 888 sites after removal of gaps, comprises the portions of two protein-coding genes (sites 1 – 694) and a transfer RNA (tRNA) region (sites 695 – 888). Transfer RNA is a highly conserved molecule in charge of translating the information encoded by coding genes into the protein alphabet. Such a translation process is achieved by mapping each set of three consecutive, non-overlapping DNA characters within a coding region into one amino acid. A coding DNA triplet is called a codon, and the second position ($cp2$) of a codon is known to undergo substitutions at slower rates than the first ($cp1$) and third codon positions ($cp3$; Fitch and Markowitz, 1970). This difference in substitution rates relates to the fact that a change at the third codon position does not always affect the resulting protein but a change at $cp2$ may, more likely, alter the final product and result in a deleterious mutation.

In this analysis, we are interested in detecting the evolutionary heterogeneity that exists between the different codon positions and the tRNA region. The primate mtDNA alignment has been analysed extensively using phylogenetic methods (Yang, 1995; Larget and Simon, 1999; Suchard et al., 2001), in most cases assuming four evolutionary classes (corresponding to the three codon positions plus the tRNA region). Most of these previous approaches have relied on prior knowledge about site membership which may be restrictive and error prone. For instance,

in a study by Yang (1995), some sites within the tRNA region were *a priori* misclassified resulting in inaccurate parameter estimates, as stated in the mtprim9.nuc file distributed with the software package PAML4 (Yang, 2007).

6.2 Model selection

We considered $Q+t$, Q and $r+t$ mixtures, with different number of components, for the primate mtDNA alignment. Figure 3 shows the log marginal likelihoods of these models, estimated using the SS sampler. The proposals were tuned to $\delta = 1.5$ for the BLM move; $\sigma = 0.06$ for BLNA; $\alpha_r = 800$, $\alpha_\pi = 600$ and $\epsilon_\theta = 0.0001$ for the ϵ Dirichlet proposal for substitution rates and stationary probabilities; and $\epsilon_\omega = 0.0001$ for mixture proportions. Hyperparameter η for the prior on a branch length was set to 2.5, in line with Suchard et al. (2001). Following a burn-in phase consisting of 20 000 iterations at power $\beta = 1$ (the posterior), the SS sampler was run for 5 000 iterations for each β value in the steppingstone path ($K = 30$ steps in total).

For comparison, we fitted the data with a discrete-gamma model using the SS sampler in MrBayes 3.2. In Figure 3 it is clear that the data contain heterogeneity that is not fully accounted for by either the homogeneous or the discrete-gamma models. A Q mixture with three components improved upon the homogeneous model by nearly 197 log-units, and upon the discrete-gamma model with 2 and 3 categories by approximately 23 and 19 log-units, respectively. In all cases, the Kass and Raftery (1995) scale indicated very strong evidence in favour of a $Q(3)$ mixture. A four-component Q mixture continued to improve upon the Q mixture with three components, but this improvement was non-significant, i.e. $2\ln(\text{BF}_{Q(4) \text{ vs } Q(3)}) < 2$. The model choice mechanism, therefore, pointed towards the $Q(3)$ mixture as the most plausible model for the data.

6.3 Model estimation and classification

Two independent runs for inference were conducted comprising 40 000 iterations of our MCMC algorithm, with no thinning, preceded by 15 000 cycles as burn in. Examination of trace plots of the log-likelihood, the observed consistency between runs and our experience with the SS runs suggested that the burn-in period was sufficiently long. We also confirmed that the runs did not

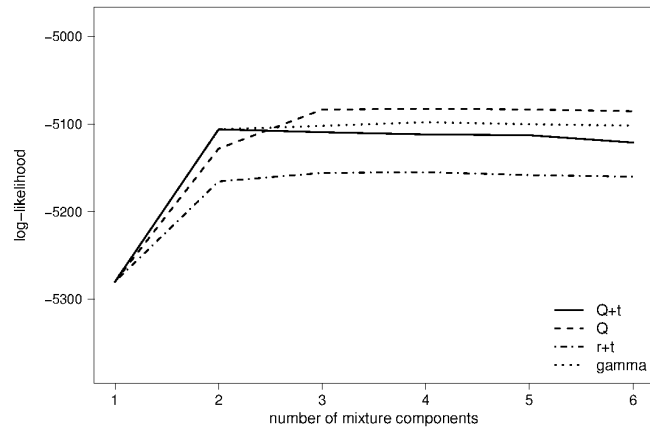


Figure 3: Estimated log marginal likelihoods for the models fitted to the primate mtDNA alignment using the SS sampler. Plotted data: homogeneous: $-5\,280.07$; $Q+t$ with 2–6 components: $-5\,106.11$, $-5\,109.42$, $-5\,111.94$, $-5\,112.85$, $-5\,121.21$; Q with 2–6 components: $-5\,128.15$, $-5\,083.56$, $-5\,082.61$, $-5\,083.46$, $-5\,085.46$; $r+t$ with 2–6 components: $-5\,165.62$, $-5\,155.88$, $-5\,155.02$, $-5\,158.40$, $-5\,160.15$; gamma with 2–6 categories: $-5\,106.20$, $-5\,102.17$, $-5\,097.97$, $-5\,100.35$, $-5\,101.87$.

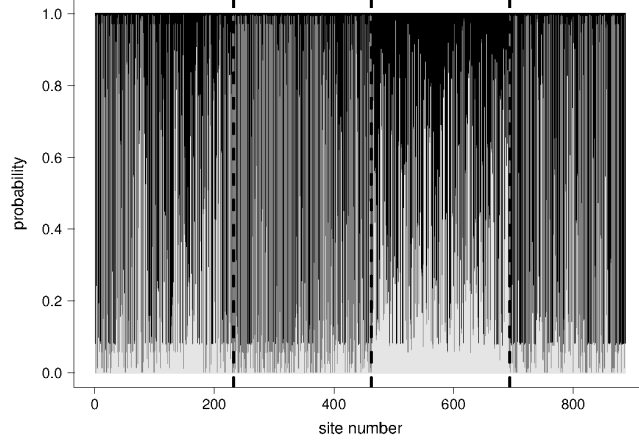


Figure 4: Posterior classification probabilities for the primate mtDNA alignment, from an analysis with a three-component Q mixture. Classification probabilities to each component are differentiated by colour: \blacksquare , \blacksquare and \blacksquare denote the three different components. The dotted lines separate the regions $cp1$ | $cp2$ | $cp3$ | tRNA in the alignment.

suffer from label-switching by examination of the MCMC trace plots. The tuning parameters for the proposals remained at the same values as for the SS run.

Figure 4 shows the estimated posterior classification probabilities of sites belonging to each of the three components in the mixture. For ease of visual interpretation, the protein-coding genes have been rearranged according to codon position; sites 1 – 232 correspond to $cp1$, sites 233 – 463 to $cp2$ and sites 464 – 694 to $cp3$, but there is nothing in the formulation of the classification method that requires such a rearrangement. Two clear patterns emerged: sites in the highly conserved $cp2$ and tRNA regions were mostly allocated to component \blacksquare , whereas the $cp3$ region is clearly dominated by components \blacksquare and \blacksquare . The method is able to capture the qualitatively different patterns of evolution in the data without prior partition into classes: the $cp1$ and $cp3$ regions are evolutionarily distinct to the $cp2$ and tRNA classes, with the bulk of this difference being observed between the $cp3$ and the $cp2$ /tRNA classes. Moreover, our approach allows for probabilistic statements of site classification such as “*site 1 belongs to component \blacksquare with probability 0.91690, to component \blacksquare with probability 0.00007 and to component \blacksquare with probability 0.08302*”.

Our approach allows inferences on evolutionary model parameters for individual mixture components and branch lengths. Table 1 reports the ergodic average of parameters for each component distribution. Components ■ and ■, which seem to specialise in *cp3*, show ergodic averages for the rates of substitution that agree with the bias that favours transitions (a substitution from $A \rightarrow G$ or $C \rightarrow T$) over transversions (any other substitution). This transition bias has been reported before (Wakeley, 1996) and one of its suggested causes is the fact that only about 3% of transitions at the third codon position cause amino acid changes compared with 41% of transversions (Crozier and Crozier, 1993). Component ■ shows extremely low *A* content, which clearly leads to poor r_{AC} , r_{AG} and r_{AT} estimates (Supplementary Material IV). This illustrates that the estimation of some parameters on an individual component basis may be discouraged for some applications for which the component-specific data do not contain the signal required to estimate all phylogenetic parameters. In such cases, one could specify mixtures of evolutionary models with constrained cases of the GTR model (Evans and Sullivan, 2012; not attempted here) or restrict inference to the allocation variables, z_1, \dots, z_N .

The consensus tree topology, obtained as the 50% majority-rule, is shown in Figure 5. This topology agrees favourably with the published topologies in Yang (1995), Larget and Simon (1999) and Suchard et al. (2001). The total length of interior and exterior branches, estimated as the ergodic average of post-burn-in samples, was 0.8574 and 2.4410, respectively (estimated uncertainties are visually reported in Supplementary Material IV).

A mixture model augmented with allocation variables was successful in describing the heterogeneity present in the primate mtDNA alignment. The analysis further allowed us to visualise underlying structural information. Regions in the alignment that are known to be highly conserved (*cp2* and *tRNA*) were grouped in a common component whereas the highly variable *cp3* region was classified to a distinct component. Such a structure discovery process could be applied to molecular sequence data for which *a priori* partition information is not available. For instance, at the moment of writing, the leading author of this paper is using phylogenetic mixture models to classify hundreds of genes into groups. The ultimate goal is to select only a few representatives per group and thus reduce the dimensionality of the problem, i.e. to move from a problem that includes hundreds of genes to one that only deals with a few.

	component ■	component ■	component ■
r_{AC}	0.0048	0.0451	0.0893
r_{AG}	0.4325	0.3981	0.2500
r_{AT}	0.0112	0.5102	0.1309
r_{CG}	0.0608	0.0039	0.0352
r_{CT}	0.4173	0.0414	0.4535
r_{GT}	0.0734	0.0013	0.0411
π_A	0.6022	0.0088	0.4269
π_C	0.2075	0.3314	0.3257
π_G	0.0259	0.1913	0.1221
π_T	0.1644	0.4685	0.1253
ω_j	0.3765	0.3858	0.2377

Table 1: Ergodic averages of model parameters from an analysis of the primate mtDNA alignment with a three-component Q mixture. For a visual measure of parameter uncertainty, refer to Supplementary Material IV.

6.4 Predictive uses for allocation variables

The allocation variables form a coherent basis for site classification. In addition to providing the means to classify the observed sites via the posterior probabilities $\{p(z_n = j \mid x); j = 1, 2, \dots, k\}$, they can also be used to classify sites on a predictive basis. To illustrate this we address the question of classifying a future (unobserved) site pattern x_n^* . If the allocation variable corresponding to such a site pattern is z_n^* and assuming k is known, then we are interested in the probability

$$p(z_n^* = j \mid x, x_n^*)$$

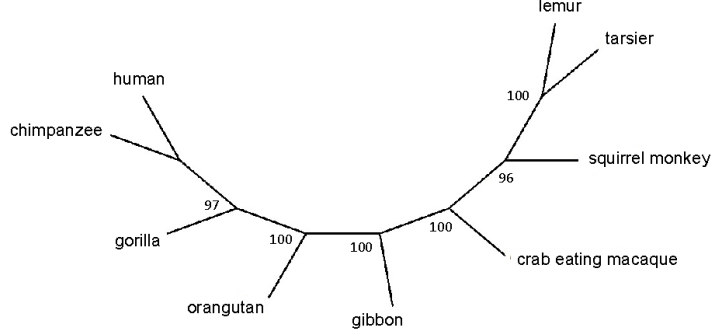


Figure 5: The 50% majority-rule consensus topology obtained from the chain of sampled topologies during the analysis of the primate mtDNA alignment with a three-component Q mixture. The numbers at the nodes indicate the percent of topologies, within the chain of sampled topologies, in which that clade is present.

which can be approximated as (Richardson and Green, 1997):

$$\begin{aligned}
p(z_n^* = j \mid x, x_n^*) &= \int p(z_n^* = j \mid x, x_n^*, \phi, t, \theta, \omega) p(\phi, t, \theta, \omega \mid x, x_n^*) d\phi dt d\theta d\omega \\
&= \int p(z_n^* = j \mid x_n^*, \phi, t, \theta, \omega) p(\phi, t, \theta, \omega \mid x, x_n^*) d\phi dt d\theta d\omega \\
&\approx \int p(z_n^* = j \mid x_n^*, \phi, t, \theta, \omega) p(\phi, t, \theta, \omega \mid x) d\phi dt d\theta d\omega
\end{aligned}$$

and the last integral is estimated by the MCMC empirical average

$$\sum_{i=1}^M \left[\omega_j p(x_n^* \mid \phi, t, \theta_j) / \sum_{j=1}^k \omega_j p(x_n^* \mid \phi, t, \theta_j) \right]_i / M \quad (19)$$

where $p(\cdot \mid \phi, t, \theta_j)$ is the j th component distribution of a Q mixture with k components and M is the number of MCMC samples after burn-in. Table 2 shows the predictive classification probabilities of 16 different site patterns, of the many tens of thousands we could consider, calculated using equation (19) and the MCMC sample generated when the $Q(3)$ mixture was

fitted to the primate mtDNA alignment (Section 6.3). All of these site patterns are unobserved in the primate alignment.

Pattern	$p(z_n^* = \blacksquare \mid \cdot)$	$p(z_n^* = \blacksquare \mid \cdot)$	$p(z_n^* = \blacksquare \mid \cdot)$
$h=C, r=A$	0.2682	0.0000	0.7318
$h=T, r=A$	0.5125	0.0000	0.4875
$h=A, r=C$	0.1069	0.0635	0.8296
$h=G, r=C$	0.1721	0.4610	0.3669
$h=T, r=C$	0.2875	0.5097	0.2028
$h=A, r=G$	0.0049	0.5320	0.4632
$h=C, r=G$	0.0016	0.8118	0.1866
$h=T, r=G$	0.0009	0.9176	0.0814
$h=A, r=T$	0.0585	0.8543	0.0872
$h=C, r=T$	0.1750	0.7638	0.0612
$h=G, r=T$	0.0909	0.8601	0.0490

Table 2: Predictive classification probabilities of sites with patterns as indicated in the first column, where h is the character in the human sequence and r is the characters in the rest of the sequences.

This exercise illustrates the potential of allocation variables as tools for classification of site patterns that are not observed in an alignment. Given an alignment of multiple molecular sequences and an MCMC run that estimates a given mixture model such as the ones presented in this paper (although not restricted to them), one could ask what is the probability of an unobserved site pattern belonging to each mixture component. This can be of direct relevance in genomic studies, where it is impractical to rerun the MCMC sampler if a new allele is sequenced; one could instead simply compute the predictive classification probabilities of those site patterns that are novel in the newly sequenced allele.

7 Discussion

We have presented a method that employs phylogenetic mixture models augmented with allocation variables to produce site classification estimates and a measure of their uncertainty.

The mixture models that we have chosen for demonstrating site classification account for both qualitative and quantitative among-site rate variation by including multiple sets of branch lengths and Q matrices. Mixtures with multiple sets of branch lengths account for a phenomenon known as heterotachy (Lopez, Casane and Philippe, 2002), in which the rates of evolution along the branches leading to different taxa in the tree vary across sites. Since the beginning of this research a number of groups have independently proposed mixtures of sets of branch lengths as a way of modelling heterotachy in phylogenetic studies (Pagel and Meade, 2008). We do note, however, that models $Q + t$, Q and $r + t$ may not be suitable in every situation. For heterotachy-free data, mixture models such as those proposed in Evans and Sullivan (2012) might suffice (i.e. mixtures of multiple evolutionary models, multiple scaling factors of these evolutionary models but only one set of branch lengths). Or, constraining Evans and Sullivan’s models one level further, the scaling factors could be made to conform to a discrete gamma distribution with empirically estimated shape, i.e., a generalisation of the discrete-gamma model of Yang (1994). This would result in less parameter-rich models than the ones presented in this work that may, or may not, provide a more plausible fit to the data in question. In problems where molecular data are suspected to have undergone recombination, a fit with either of $Q + t$, Q or $r + t$ could be inadequate because contiguous subsequences of recombinant bacterial or viral DNA would be expected to follow different phylogenetic histories; a model with class-specific topologies would then be required (Kitchen et al., 2009). Whatever the flavour of the phylogenetic mixture, we hope that the main message of our paper is clear: Bayesian mixture models can be extended to include allocation variables and be readily used as tools for site classification. The reader is encouraged to always follow the principles of a valid inferential process by considering a set of candidate models and conducting model selection before estimating a model (Fisher, 1922). It is clear that more flexible and user-friendly software tools to conduct such model selection processes in phylogenetics are required.

A potential application of our classification method, not pursued here, is as a tool for iden-

tifying the sites that are unable to undergo substitution. The presence of invariant sites is a well-documented cause of inconsistency in phylogeny reconstruction (Steel et al., 2000), and ‘augmented’ phylogenetic mixtures could be deployed to pinpoint the invariant sites that should be excluded from the alignment before inference. This idea has been discussed elsewhere (Huelsenbeck and Suchard, 2007) and its implementation would require defining a mixture that includes a strictly invariant class (i.e. a component in which all the rates of substitution are zero).

We recognise the limitations of the NNI proposal in our MCMC sampler and note that larger problems may require additional topology proposal mechanisms. The topology update mechanism that our models require is restricted to updating the tree topology while preserving the branch lengths due to the fact that there is only one topology shared across all components. The applications that we present in this paper are of modest size and we are confident that NNI successfully leads the MCMC chain into stationarity; visual inspection of the post-burn-in log-likelihood trace of the primate mtDNA alignment indicate lack of pre-stationary trends and our 50% majority-rule consensus tree agrees favourably with several other published studies. Tackling larger problems may require additional MCMC mechanisms and this is regarded as an aspect for future work.

Acknowledgements

Elisa Loza was supported by the Mexican Council for Science and Technology (CONACYT) during the production of the research and partially supported by Rothamsted Research during the preparation of the manuscript. Rothamsted Research receives grant-aided support from the Biotechnology and Biological Sciences Research Council (BBSRC) of the UK. Merrilee Hurn and Tony Robinson were supported by the Bath Institute for Complex Systems (EPSRC grant GR/S86525/01). We thank Klaus Kurtenbach, Gabi Margos and Ed Feil, all from the Department of Biology and Biochemistry at the University of Bath, for their helpful discussions. Thanks also to Kevin Dawson, from the Wellcome Trust Sanger Institute, for valuable comments to earlier versions of this manuscript. We thank two anonymous reviewers for providing valuable criticism that improved the quality of this paper.

References

- [1] ARDIA, D., BAŞTÜRK, N., HOOPERHEIDE, L., AND VAN DIJK, H. K. A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood. *Computational Statistics and Data Analysis* 56 (2012), 3398–3414.
- [2] BAELE, G., LEMEY, P., BEDFORD, T., RAMBAUT, A., SUCHARD, M. A., AND ALEKSEYENKO, A. V. Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty. *Molecular Biology and Evolution* 29, 9 (2012), 2157–2167.
- [3] BROWN, W. M., PRAGER, E. M., WANG, A., AND WILSON, A. C. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *Journal of Molecular Evolution* 18 (1982), 225–239.
- [4] CALDERHEAD, B., AND GIROLAMI, M. Estimating Bayes factors via thermodynamic integration and population MCMC. *Computational Statistics and Data Analysis* 53 (2009), 4028–4045.
- [5] CROZIER, R. H., AND CROZIER, Y. C. The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organization. *Genetics* 133 (1993), 97–117.
- [6] EVANS, J., AND SULLIVAN, J. Generalized Mixture Models for Molecular Phylogenetic Estimation. *Systematic Biology* 61, 1 (2012), 12–21.
- [7] FAN, Y., WU, R., CHEN, M.-H., KUO, L., AND LEWIS, P. O. Choosing among Partition Models in Bayesian Phylogenetics. *Molecular Biology and Evolution* 28, 1 (2011), 523–432.
- [8] FELSENSTEIN, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* 17 (1981), 368–376.
- [9] FELSENSTEIN, J., AND CHURCHILL, G. A. A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution* 13, 1 (1996), 93–104.
- [10] FISHER, R. A. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society (Series A)* 222 (1922), 309–368.

- [11] FITCH, W. M., AND MARKOWITZ, E. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics* 4 (1970), 579–593.
- [12] FRIEL, N., HURN, M., AND WYSE, J. Improving power posterior estimation of statistical evidence. To appear in *Statistics and Computing*, DOI: 10.1007/s11222-013-9397-1, 2013.
- [13] FRIEL, N., AND PETTITT, A. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society B* 70, 3 (2008), 589–607.
- [14] GREEN, P. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82, 4 (1995), 711–732.
- [15] HASEGAWA, M., KISHINO, H., AND YANO, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22 (1985), 160–174.
- [16] HAYASAKA, K., GOJOBORI, T., AND HORAI, S. Molecular phylogeny and evolution of primate mitochondrial DNA. *Molecular Biology and Evolution* 5, 6 (1988), 626–644.
- [17] HUELSENBECK, J., AND SUCHARD, M. A Nonparametric Method for Accommodating and Testing Across-Site Rate Variation. *Systematic Biology* 56, 6 (2007), 975–987.
- [18] HUELSENBECK, J. P., AND RONQUIST, F. MrBayes: Bayesian inference of phylogeny. *Bioinformatics* 17 (2001), 754–755.
- [19] HURN, M. A., GREEN, P. J., AND AL-AWADHI, F. A Bayesian hierarchical model for photometric redshifts. *Applied Statistics, Journal of the Royal Statistical Society C* 57 (2008), 487–504.
- [20] JASRA, A., HOLMES, C. C., AND STEPHENS, D. A. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science* 20, 1 (2005), 50–67.
- [21] JUKES, T., AND CANTOR, C. Evolution of protein molecules. In *Mammalian Protein Metabolism*, H. N. Munro, Ed. Academic Press, New York, USA, 1969, pp. 21–132.

- [22] KASS, R. E., AND RAFTERY, A. E. Bayes Factors. *Journal of the American Statistical Association* 90, 430 (1995), 773–795.
- [23] KITCHEN, C. M. R., KROLL, J., KURITZKES, D. R., BLOOMQUIST, E., DEEKS, S. G., AND SUCHARD, M. A. Two-way Bayesian hierarchical phylogenetic models: An application to the co-evolution of gp120 and gp41 during and after enfuvirtide treatment. *Computational Statistics and Data Analysis* 53 (2009), 766–775.
- [24] KOLACZKOWSKI, B., AND THORNTON, J. W. A mixed branch length model of heterotachy improves phylogenetic accuracy. *Molecular Biology and Evolution* 25, 6 (JUN 2008), 1054–1066.
- [25] LANAVE, C., PREPARATA, G., SACCONI, C., AND SERIO, G. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution* 20 (1984), 86–93.
- [26] LARGET, B., AND SIMON, D. L. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution* 16, 6 (1999), 750–759.
- [27] LARTILLOT, N., AND PHILIPPE, H. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Molecular Biology and Evolution* 21, 6 (2004), 1095–1109.
- [28] LARTILLOT, N., AND PHILIPPE, H. Computing Bayes factors using thermodynamic integration. *Systematic Biology* 55 (2006), 195–207.
- [29] LESLIE, D. S. Discussion on Model-based clustering for social networks (by M. S. Handcock, A. E. Raftery and J. M. Tantrum). *Journal of the Royal Statistical Society A* 170 (2007), 301–354.
- [30] LOPEZ, P., CASANE, D., AND PHILIPPE, H. Heterotachy, an important process of protein evolution. *Molecular Biology and Evolution* 19, 1 (JAN 2002), 1–7.
- [31] MEADE, A., AND PAGEL, M. A Phylogenetic Mixture Model for Heterotachy. In *Evolutionary Biology from Concept to Application*, P. Pontarotti, Ed., first ed. Springer-Verlag, 2008, pp. 29–41.

- [32] MOORE, G. W., GOODMAN, M., AND BARNABAS, J. An iterative approach from the stand-point of the additive hypothesis to the dendrogram problem posed by molecular data sets. *Journal of Theoretical Biology* 38 (1973), 423–457.
- [33] NEWTON, M., AND RAFTERY, A. Approximate Bayesian inference by the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series B* 56 (1994), 3–48.
- [34] PAGEL, M., AND MEADE, A. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology* 56, 4 (2004), 571–581.
- [35] PAGEL, M., AND MEADE, A. Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. *Philosophical Transactions of the Royal Society B* 363 (2008), 3955–3964.
- [36] RAMBAUT, A., AND GRASSLY, N. C. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences* 13 (1997), 235–238.
- [37] RICHARDSON, S., AND GREEN, P. J. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society B* 59, 4 (1997), 731–758.
- [38] ROBINSON, D. F. Comparison of labeled trees with valency three. *Journal of Combinatorial Theory* 11 (1971), 105–119.
- [39] RONQUIST, F., TESLENKO, M., VAN DER MARK, P., AYRES, D. L., DARLING, A., HOHNA, S., LARGET, B., LIU, L., SUCHARD, M. A., AND HUELSENBECK, J. P. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology* 61, 3 (2012), 539–542.
- [40] STEEL, M. Should phylogenetic models be trying to "fit an elephant". *Trends in Genetics* 21, 6 (2005), 307–309.
- [41] STEEL, M., HUSON, D., AND LOCKHART, P. J. Invariable Sites Models and Their Use in Phylogeny Reconstruction. *Systematic Biology* 49 (2000), 225–232.

- [42] SUCHARD, M. A., WEISS, R. E., AND SINSHEIMER, J. S. Bayesian selection of continuous-time Markov chain evolutionary models. *Molecular Biology and Evolution* 18, 6 (2001), 1001–1013.
- [43] TAVARÉ, S. Some probabilistic and statistical problems in the analysis of DNA sequences. In *Lectures on Mathematics in the Life Sciences*, R. M. Miura, Ed., vol. 17. American Mathematical Society, Providence, USA, 1986, pp. 57–86.
- [44] VERDINELLI, I., AND WASSERMAN, L. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association* 90 (1995), 614–618.
- [45] WAKELEY, J. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends in Ecology and Evolution* 11, 4 (1996), 158–163.
- [46] WEBB, A., HANCOCK, J. M., AND HOLMES, C. C. Phylogenetic inference under recombination using Bayesian stochastic topology selection. *Bioinformatics* 25, 2 (2009), 197–203.
- [47] XI, Z., RUHFEL, B. R., SCHAEFER, H., AMORIM, A. M., SUGUMARAN, M., WURDACK, K. J., ENDRESS, P. K., MATTHEWS, M. L., STEVENS, P. F., MATHEWS, S., AND DAVIS, C. C. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *PNAS* 109, 43 (2012), 17519–17524.
- [48] XIE, W., LEWIS, P. O., FAN, Y., AND CHEN, M.-H. Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection. *Systematic Biology* 60 (2011), 150–160.
- [49] YANG, Z. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* 10, 6 (1993), 1396–1401.
- [50] YANG, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* 39 (1994), 306–314.

- [51] YANG, Z. A space-time process model for the evolution of DNA sequences. *Genetics* 139 (1995), 993–1005.
- [52] YANG, Z. *Computational Molecular Evolution*. Oxford Series in Ecology and Evolution. Oxford University Press, Great Britain, 2006.
- [53] YANG, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* 24 (2007), 1586–1591.